DOCUMENT RESUME

ED 479 596 IR 022 200

AUTHOR	Person, Natalie; Erkel, Mary; Graesser, Arthur C.
TITLE	AutoTutor Passes the Bystander Turing Test.
PUB DATE	2002-10-00
NOTE	6p.; In: E-Learn 2002 World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education. Proceedings (7th, Montreal, Quebec, Canada, October 15-19, 2002); see IR 022 093.
AVAILABLE FROM	Association for the Advancement of Computing in Education (AACE), P.O. Box 3728, Norfolk, VA 23514. Tel: 757-623-7588; e-mail: info@aace.org; Web site: http://www.aace.org/DL/.
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS
DESCRIPTORS	Animation; Computer Assisted Instruction; *Computer Simulation; Computer System Design; Educational Technology; Instructional Effectiveness; *Intelligent Tutoring Systems; *Programmed Tutoring; *Tutoring; Tutors

ABSTRACT

In this study, the Bystander Turing Test (BTT) paradigm is used to determine whether participants rate particular dialog moves in tutoring transcripts to be generated by a computer tutor, AutoTutor, or by a skilled human tutor. Currently, AutoTutor initiates a conversation with the learner by posing a question or presenting a problem for the learner to solve. Students learn about computer literacy or physics by engaging in a conversation with an animated agent that remains on the monitor throughout the tutoring session. AutuTutor scaffolds the conversation with a series of dialog moves that are frequently used by effective human tutors. Results indicate that participants are unable to differentiate computer-generated moves from those generated by skilled human tutors. Participants did, however, view some dialog move categories to be more pedagogically effective than others regardless of whether the dialog move was generated by AutoTutor or by a skilled human tutor. (Contains 17 references.) (Author)

Reproductions supplied by EDRS are the best that can be made
from the original document.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

G.H. Marks

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

AutoTutor Passes the Bystander Turing Test

Natalie Person[1], Mary Erkel[1], Arthur C. Graesser[2],
& the Tutoring Research Group[2]
[1] Rhodes College, Department of Psychology, Memphis, TN 38112
person@rhodes.edu
[2] University of Memphis, Department of Psychology, Memphis, TN 38152-3230
a-graesser@memphis.edu

ED 479 596

Abstract: In this study, the Bystander Turing Test (BTT) paradigm is used to determine whether participants rate particular dialog moves in tutoring transcripts to be generated by a computer tutor, AutoTutor, or by a skilled human tutor. Results indicate that participants are unable to differentiate computer-generated moves from those generated by skilled human tutors. Participants did, however, view some dialog move categories to be more pedagogically effective than others regardless of whether the dialog move was generated by AutoTutor or by a skilled human tutor.

Introduction

The study presented in this paper is one of several studies that were recently conducted to determine whether AutoTutor can pass a variation of the Turing test. AutoTutor is a web-based computer tutor architecture that can be used for a variety of content domains (Graesser, Hu, Susarla, Harter, Person, Louwerse, Olde, & the TRG, 2001; Graesser, Person, Harter, & the TRG, 2000; Person, Graesser, Bautista, Mathews, & the TRG, 2001; Person, Graesser, Kreuz, Pomeroy, & the TRG, 2001). AutoTutor's architecture has been described extensively in previous publications, and therefore, will not be addressed here (Graesser, Wiemer-Hastings, Wiemer-Hastings, Harter, Person, & the TRG, 2000; Person, Graesser, & the TRG, 2000; Person, Graesser, Harter, Mathews, & the TRG, 2000; Person et al., 2001; Person, Klettke, Link, Kreuz, & the TRG, 1999; Wiemer-Hastings, Graesser, Harter, & the TRG, 1998; Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999). The Tutoring Research Group (TRG) has recently developed two versions of AutoTutor, one for computer literacy and one for conceptual physics. The computer literacy AutoTutor is designed to help students learn basic computer literacy topics covered in an introductory course (e.g., hardware, operating systems, and the Internet). The conceptual physics AutoTutor is designed to help students learn Newtonian physics.

Two of the primary design goals of AutoTutor's developers are to create a system that simulates the dialog moves of effective human tutors and to have the system deliver the dialog moves in conversationally appropriate ways. We know that effective teachers and tutors encourage students to take an active role in the learning process by structuring tasks and dialogs in ways that force students to engage in behaviors such as asking questions, recognizing misconceptions, generating explanations, and synthesizing information from multiple sources (Chi, 1996; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Graesser et al., 2001; Person et al., 2001). Currently AutoTutor initiates a conversation with the learner by posing a question or presenting a problem for the learner to solve. Students learn about computer literacy or physics by engaging in a conversation with an animated agent that remains on the monitor throughout the tutoring session. AutoTutor scaffolds the conversation with a series of dialog moves that are frequently used by effective human tutors (Graesser et al., 2001; Person & Graesser, 1999). The dialog moves included in AutoTutor are Pump, Prompt, Hint, Assertion, Correction, Summary, and three kinds of Short Feedback (positive, negative, and neutral). Although AutoTutor delivers these dialog moves (via the animated agent) in pedagogically and conversationally appropriate ways that enable learners to construct elaborate answers and solutions to questions and problems (Person et al., 2001), it is unclear whether AutoTutor is actually simulating the pedagogical and conversational behaviors of human tutors or whether AutoTutor is producing a set idiosyncratic behaviors that accomplish the same end. The present study is designed to elucidate whether participants view the teaching and conversational behaviors of AutoTutor differently from those of skilled human tutors. The subsequent sections of this paper include the following: (1) an overview of the Bystander Turing Test (BTT), (2) a description of the study being presented in this paper, (3) a summary of the findings and (4) the conclusions drawn from the BBT research.

IR022200

BEST COPY AVAILABLE

The Bystander Turing Test

Turing (1950) argued that computers are not capable of thinking and provided several theological, psychological, and sociological arguments in support his position. To determine a computer program's intelligence, Turing proposed several benchmark methods. One such method requires humans to decide whether they are interacting with an actual computer program or another human via computer mediation. According to Turing, a computer could be described as intelligent if it could deceive a human into believing that it was human.

The original Turing test is based on the Imitation Game. In the Imitation Game, the participants are a man, a woman, and an interrogator. The interrogator is physically separated from the man and woman. The object of the game is for the interrogator to determine which participant is male and which is female by evaluating their responses to questions posed by the interrogator. Turing proposed replacing one of the humans (i.e., the man or woman) with a computer. If the interrogator cannot discern whether the responses are generated by the computer or by the human, the computer program is said to pass the Turing Test. Specifically, the computer program is emulating human thought and intelligence if it can process human language, utilize knowledge it has been given before and receives during the interactive session, and use stored knowledge to engage in an intellectual conversation with the learner (Turing, 1950).

The Bystander Turing Test (BTT) is a variation of the original Turing Test (Person, Graesser, & the TRG, 2002). In the BTT, participants rate whether particular dialog moves in tutoring transcripts are generated by AutoTutor or by skilled human tutors. We converged on this variation of the original Turing Test for two reasons that are based on our observations of human tutors attempting to tutor students in computer-mediated environments. First, human tutors have tremendous difficulty responding in conversational real-time to student input. Given that AutoTutor responds immediately to student contributions, noticeable time lags resulting from human tutors' need to think would potentially influence an interrogator's decisions. Second, human tutors frequently make errors when typing their responses. Hence, AutoTutor's text-to-speech generator would mispronounce a considerable number of words, and unquestionably, affect an interrogator's judgments.

Recently Person, Graesser, and the TRG (2002) used the BTT paradigm in a series of studies to determine the following: (1) whether participants could discriminate dialog moves that were generated by skilled human tutors from those generated by AutoTutor, (2) whether the pedagogical effectiveness of the dialog moves generated by AutoTutor differed from the those generated by human tutors, and (3) whether the participants perceived the dialog generated by AutoTutor to be more or less conversationally appropriate than those generated by humans. The results of these studies indicated that the participants could not discriminate whether particular dialog moves were generated by AutoTutor or by human tutors. Further, they reported no mean differences on measures of pedagogical effectiveness and conversational appropriateness for AutoTutor and human-generated dialog moves.

One explanation of why the speaker discrimination (computer or human) was so low involves the dialog move categories. Perhaps the college students used the dialog move category information to diagnostically decide whether a computer or human produced the dialog move. College students may have been penalized in their predictions if they were poor guessers for some categories, but accurate guessers for others. However, previous findings from a dialog move category analysis indicated that the likelihood that the participants believed that AutoTutor generated a dialog move was absolutely equivalent for all of the dialog move categories. That is, the proportion of four, five, and six ratings on the "Who Said It?" question for each dialog move category did not significantly differ (Person, Graesser, & the TRG, 2002).

One interpretation of the findings from the previous BBT studies is that participants were simply not able to discriminate anything (e.g., speaker, examples of effective teaching, conversational appropriateness, or dialog move categories), and therefore, were forced to guess on every rating dimension. The study presented in this paper addresses this very issue by including randomly inserted dialog moves in the conversations. It is predicted that the participants will be sensitive to randomly inserted tutor dialog moves in that they will perceive the random moves to be generated by the computer (i.e., AutoTutor). It is further predicted that the random moves will not be rated as favorably on the Effective Teaching and Conversational Appropriateness assessment questions.

The BTT Study with Random Dialog Moves

Participants. The participants were 64 undergraduates at a small liberal arts college located in Memphis, Tennessee. The students participated in the study to fulfill research requirements for introductory psychology classes.

Materials. The conversations included in this BTT study were constructed in the following way. Two hundred and eighty-two conversations were randomly selected from several thousand conversations that had taken place between college students and the computer literacy AutoTutor. In each of the 282 conversations, a particular AutoTutor dialog move was deleted along with all subsequent conversational turns. Six skilled human computer literacy tutors were asked to read each conversation and then fill in the blank lines with what they would say to that student at that juncture in the conversation.

After the human tutor responses (i.e., dialog moves) were collected, packets containing 36 different tutoring conversations were assembled. Each packet of 36 contained four types of conversations: (1) conversations in which all tutor dialog moves were generated by AutoTutor in actual tutoring sessions with students (Authentic computer condition, 9 total), (2) conversations in which the last tutor turn was a randomly selected AutoTutor dialog move from another tutoring conversation on the same topic (Random Computer condition, 9 total), (3) conversations in which the last dialog move in the conversation was generated by a human tutor (Authentic Human condition, 9 total), and (4) conversations in which the last tutor turn was a randomly selected human dialog move from another tutoring conversation on the same topic (Random Human condition, 9 total). To the extent possible, the dialog move categories (e.g., Pump, Prompt, Assertion) for both AutoTutor and the humans were evenly distributed across the four conversation types.

Procedure. All participants read and signed an informed consent form before the experiment began. Testing packets were then given to the participants that included 36 conversations. After reading each conversation, participants provided three six-point scale ratings for the last tutor dialog move in each conversation (see Table 1). After providing ratings for all 36 conversations, participants were asked to

Table 1. Dialog Move Assessment Questions Used in Study 1 and 2

1. The last tutor response/question is:

1	2	3	4	5	6
Definitely Human	Probably Human	Not sure, but guess Human	Not sure, but guess Human	Probably Computer	Definitely Computer

2. In terms of conversational appropriateness, the last tutor response/question is:

1	2	3	4	5	6
Definitely Inappropriate	Probably Inappropriate	Not sure, but guess Inappropriate	Not sure, but guess Appropriate	Probably Appropriate	Definitely Appropriate

3. In terms of effective teaching, the last tutor turn response/question is:

Definitely Ineffective	Probably Ineffective	Not sure, but guess Ineffective	Not sure, but guess Effective	Probably Effective	Definitely Effective

describe any strategies they had developed for identifying which dialog moves were generated by the computer and which were generated by humans. The analysis of these open-ended responses will not be discussed in this paper.

Results

Three 2 x 2 ANOVAs were performed to determine the effects of Speaker (Computer vs. Human) and Authenticity (Authentic vs. Random) on the participants' ratings of the three six-point scale assessment questions. The means for the three questions are reported in Table 2.

Table 2. Means and Standard Deviations for the Three Assessment Questions

Question	Computer		Human		Total	
	Mean (n)	SE	Mean (n)	SE	Mean (n)	SE
1. Who said it?						
Authentic	3.62 (36)	0.12	3.59 (36)	0.13	3.60 (72)	0.09
Random	3.91 (36)	0.12	3.88 (36)	0.12	3.90 (72)	0.09
Total	3.77 (72)	0.09	3.73 (72)	0.09		

2. Conversationally appropriate?			
Authentic	4.24 (36) 0.14	4.41 (36) 0.14	4.33 (72) 0.10
Random	3.45 (36) 0.14	3.25 (36) 0.14	3.35 (72) 0.10
Total	3.85 (72) 0.10	3.83 (72) 0.10	
3. Effective teaching?			
Authentic	4.00 (36) 0.13	4.09 (36) 0.14	4.04 (72) 0.10
Random	3.26 (36) 0.14	3.29 (36) 0.14	3.28 (72) 0.10
Total	3.63 (72) 0.10	3.69 (72) 0.10	

Who said it? (Question 1). The 2 x 2 ANOVA for the "Who said it?" question indicated no mean differences for the Speaker condition, $F(1,140) = 0.07$, $p > .05$. The participants could not distinguish the dialog moves that were generated by AutoTutor (Mean = 3.77) from those that were generated by skilled human tutors (Mean = 3.73). However, the means for the Authenticity condition did significantly differ, $F(1,140) = 5.63$, $p < .05$. Participants considered the Random dialog moves (Mean = 3.90) to be significantly more "computer-like" than the Authentic dialog moves (Mean = 3.60). The interaction between Speaker and Authenticity was not statistically significant, $F(1,140) = 0.00$, $p > .05$.

Conversationally Appropriate? (Question 2). The 2 x 2 ANOVA for the "Conversational Appropriate?" question indicated no mean differences for the Speaker condition, $F(1,140) = 0.02$, $p > .05$. The participants considered the dialog moves generated by AutoTutor (Mean = 3.85) to be as conversationally appropriate as those generated by the human tutors (Mean = 3.83). The means for the Authenticity condition, however, did significantly differ, $F(1,140) = 49.65$, $p < .05$. As predicted, the Authentic dialog moves (Mean = 4.33) were rated as more conversationally appropriate than the Random dialog moves (Mean = 3.35). The interaction between the two conditions was not statistically significant, $F(1,140) = 1.83$, $p > .05$.

Effective Teaching? (Question 3). The 2 x 2 ANOVA for the "Effective Teaching?" question indicated no mean differences for the Speaker condition, $F(1,140) = 0.18$, $p > .05$. The Effective Teaching ratings for the dialogs moves generated by AutoTutor (Mean = 3.63) were equal to those generated by the human tutors (Mean = 3.69). Once again, the means for the Authenticity condition did significantly differ, $F(1,140) = 31.56$, $p < .05$. The Authentic dialog moves (Mean = 4.04) received higher effective teaching ratings than the Random moves (Mean = 3.28). The interaction term for the two conditions was not statistically significant, $F(1,140) = 0.04$, $p > .05$.

Discussion

We can conclude from this study that AutoTutor is achieving many of the design goals of AutoTutor's developers by simulating the dialog moves of effective human tutors and delivering them in conversationally appropriate ways. The findings from this study bolster previously reported BBT results with the addition of the Authenticity conditions. Given that the participants rated random dialog moves less favorably than authentic ones, but could not discriminate the speaker (human or computer), we can be more confident that AutoTutor's dialog move selections are on par with those of human tutors.

The fact that the Conversationally Appropriate means did not differ for speakers and were higher for authentic dialog moves is a promising indicator that AutoTutor is a competent conversational partner. Additional evidence of AutoTutor's communicative competence lies in the fact the conversational ratings for the AutoTutor authentic moves were closer to the more favorable extreme of the six-point scale. Of course, it may be the case that the human tutors adopted the conversational style of AutoTutor rather than generating the dialog moves they would have produced in an actual human-to-human tutoring session. However, even if this was the case, the participants considered the dialog moves of both speakers to be conversationally appropriate. Hence, even if AutoTutor does not emulate the style of human tutors in human-to-human sessions, AutoTutor does deliver dialog moves that are perceived as acceptable and believable.

References

Chi, M.T.H. (1996). Constructing self explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology, 10*, S33-S49.

Chi, M.T.H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989) Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.

Graesser, A.C., Hu, X., Susarla, S., Harter, D., Person, N.K., Louwerse, M., Olde, B., & the Tutoring Research Group (2001). AutoTutor: An intelligent tutor and conversational tutoring scaffold. *Proceedings for the 10th International Conference of Artificial Intelligence in Education* (pp. 47-49). San Antonio, TX.

Graesser, A.C., Person, N.K., & Magliano, J.P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring sessions. *Applied Cognitive Psychology, 9*, 1-28.

Graesser, A.C., Person, N.K., Harter, D., & the Tutoring Research Group (2000). Tactics in tutoring in AutoTutor. In the *ITS 2000 Proceedings of the Workshop on Modeling Human Teaching Tactics and Strategies* (pp. 49-57), Montreal, Canada.

Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & TRG (1999). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, 1, 35-51.

Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & the Tutoring Research Group (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments, 8*, 129-148.

Person, N.K., & Graesser, A.C. (1999). Evolution of discourse in cross-age tutoring. In A. M.O'Donnell and A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 69-86), Mahwah, NJ: Erlbaum.

Person, N. K., Graesser, A. C., & the Tutoring Research (2002). Group Human or computer: AutoTutor in a bystander Turing test. In the *Proceedings of ITS 2002*, Berlin: Springer-Verlag.

Person, N.K., Graesser, A.C., & the Tutoring Research Group (2000). Designing AutoTutor to be an effective conversational partner. In the *Proceedings for the 4th International Conference of the Learning Sciences* (pp. 246-253), Ann Arbor, MI.

Person, N.K., Graesser, A.C., Bautista, L., Mathews, E.C., & the Tutoring Research Group (2001). Evaluating Student Learning Gains in Two Versions of AutoTutor. In J.D. Moore, C.L. Redfield, & W.L. Johnson (Eds.) *Artificial intelligence in education: AI-ED in the wired and wireless future* (pp. 286-293), Amsterdam, IOS Press.

Person, N.K., Graesser, A.C., Harter, D., Mathews, E.C., & the Tutoring Research Group (2000). Dialog move generation and conversation management in AutoTutor. *Proceedings of the AAAI Fall Symposium: Building Dialogue Systems for Tutorial Applications* (pp. 45-51), Falmouth, MA: AAAI Press.

Person, N.K., Graesser, A.C., Kreuz, R.J., Pomeroy, V., & the Tutoring Research Group (2001). Simulating human tutor dialog moves in AutoTutor. *International Journal of Artificial Intelligence in Education, 12*, 23-29.

Person, N.K., Klettke, B., Link, K., Kreuz, R.J., & the Tutoring Research Group (1999). The integration of affective responses into AutoTutor. *Proceedings of the International Workshop on Affect in Interactions* (pp. 167-178), Siena, Italy.

Turing, A.M. (1950). Computing Machinery and Intelligence. In E.A. Feigenbaum & J. Feldman (Eds.) *Computers and thought*. New York: McGraw-Hill.

Wiemer-Hastings, P., Graesser, A.C., Harter, D., and the Tutoring Research Group (1998). The foundations and architecture of AutoTutor. *Proceedings of the 4th International Conference on Intelligent Tutoring Systems* (pp.334-343), Berlin, Germany: Springer-Verlag.

Wiemer-Hastings, P., Wiemer-Hastings, K., and Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. *Artificial Intelligence in Education* Amsterdam (pp. 535-542), IOS Press.

Acknowledgements

This research was supported by grants from the Office of Naval Research (N00014-00-1-0600) and the National Science Foundation (SBR 9720314 and SBR 0106965). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR or NSF.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").